Genetic Forum 24/4/2008

MRC Epidemiology Unit

Haplotype analysis

Jing Hua Zhao

Outline

- Background
- Chromosome X analysis
- Power calculation
- Associate issues

Start with a Problem …

- 5HT2 and Schizophrenia
- HLA markers and Schizophrenia
- Further work from an early report in Lancet
- Using unrelated individuals rather than the popular TDT
- But
  - Too many analyses
  - Available program (EH) would not work

Methodological Issues

Human Heredity

Hum Hered 2000;50:133–139

Received: December 17, 1998
Revision received: February 15, 1999
Accepted: February 23, 1999

Model-Free Analysis and Permutation Tests for Allelic Associations

Jing Hua Zhao[a] David Curtis[b] Pak Chung Sham[a]

[a]Department of Psychological Medicine, Institute of Psychiatry, and [b]Joint Academic Department of Psychological Medicine, St. Bartholomew's and Royal London School of Medicine and Dentistry, London, UK

Advances and Problems …

- Advances
  - Easier analysis of gene-trait association
  - Model-free statistics
  - Permutation tests
  - Some haplotype-specific statistics (F-T)
- Problems
  - Too slow
  - Potentially useful model-based measure of LD

Human Molecular Genetics, 2000, Vol. 9, No. 20 2993–2999

Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the *ALDH2* locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb

H.G. Koch, J. McClay, E.-W. Loh, S. Higuchi[1], J.-H. Zhao, P. Sham, D. Ball and I. W. Craig+

SGDP Research Centre, Institute of Psychiatry, Denmark Hill, London SE5 8AF, UK and [1]Kurihama National Hospital, Kanagawa, Japan

Original Paper

Human Heredity

Hum Hered 2002;53:36–41

Received: September 10, 2001
Revision received: December 6, 2001
Accepted: December 10, 2001

Faster Haplotype Frequency Estimation Using Unrelated Subjects

Jing Hua Zhao [a] Pak Chung Sham [a, b, c]

[a] Department of Psychological Medicine, [b] Department of Biostatistics and Computing, [c] Social, Genetic, Developmental Psychiatry Research Centre, Institute of Psychiatry, London, UK

Advances and Problems…

- Advances
  - Faster computation
  - Likelihood-based LD statistics
- Problems
  - Abandon model-specific statistics
  - Do not handle missing data
  - Still use global association test
  - Do not handle covariates

BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 12 2002
Pages 1694–1695

GENECOUNTING: haplotype analysis with missing genotypes

Jing Hua Zhao[1,], Sebastien Lissarrague[2], Laurent Essioux[3] and Pak Chung Sham[4]*

[1]Department of Epidemiology and Public Health, University College London, 1–19 Torrington Place, London WC1E 6BT, UK, [2]Genset SA, Site SNECMA RN7, 91030 Evry, France, [3]ValiGen SA, Tour Neptune, 92086 Paris-La-Défense, France and [4]Section of Genetic Epidemiology, PO Box 80, Institute of Psychiatry, London SE 8AF, UK

Received on November 22, 2001; revised on March 25, 2001; accepted on May 21, 2001

Advances and Problems

- Advances
  - Provide generic algorithm for missing genotype data as well as likelihood estimation and handle multiallelic markers
  - Some haplotype specific tests
  - Considerable easier than HAPLO (Yale)
- Problems
  - Chromosome X data
  - Slow for large problem

PLOS BIOLOGY

GAD2 on Chromosome 10p12 Is a Candidate Gene for Human Obesity

Philippe Boutin[1☯], Christian Dina[1☯], Francis Vasseur[1,2☯], Séverine Dubois[1¤1], Laetitia Corset[1], Karin Séron[1], Lynn Bekris[3], Janice Cabellon[3], Bernadette Neve[1], Valérie Vasseur-Delannoy[1], Mohamed Chikri[1¤2], M. Aline Charles[4], Karine Clement[5], Ake Lernmark[3], Philippe Froguel[1,6*]

1 Institute of Biology–Centre National de la Recherche Scientifique, Pasteur Institute, Lille, France, 2 University Hospital of Lille, Lille, France, 3 Department of Medicine, University of Washington, Seattle, Washington, United States of America, 4 Institut National de la Santé et de la Recherche Médicale (INSERM), Paul Brousse Hospital, Villejuif, France, 5 Paris VI University and INSERM "Avenir," Department of Nutrition, Hôtel Dieu Hospital, Paris, France, 6 Hammersmith Genome Centre and Department of Genomic Medicine, Imperial College, London, United Kingdom

European Journal of Human Genetics (2005) 13, 849–855

www.nature.com/ejhg

ARTICLE

Haplotype construction of the FRDA gene and evaluation of its role in type II diabetes

Johan Holmkvist*[,1], Peter Almgren[1], Hemang Parikh[1], Marco Zucchelli[2], Juha Kere[2,3,4], Leif Groop[1] and Cecilia M Lindgren[2,3]

[1]*Department of Clinical Sciences – Diabetes and Endocrinology, Malmö University Hospital, Lund University, S-205 02 Malmö, Sweden;* [2]*Department of Biosciences at Novum, Karolinska Institutet, S-14157 Huddinge, Sweden;* [3]*Clinical Research Centre, Karolinska University Hospital at Huddinge, Sweden;* [4]*Department of Medical Genetics, University of Helsinki, 00014 Helsinki, Finland*

OPEN ACCESS Freely available online

PLOS CLINICAL TRIALS

Genetic Polymorphisms and Weight Loss in Obesity: A Randomised Trial of Hypo-Energetic High- versus Low-Fat Diets

Thorkild I. A. Sørensen[1*], Philippe Boutin[2], Moira A. Taylor[3], Lesli H. Larsen[4], Camilla Verdich[1], Liselotte Petersen[1], Claus Holst[1], Søren M. Echwald[4], Christian Dina[2], Søren Toubro[5], Martin Petersen[5], Jan Polak[6], Karine Clément[7], J. Alfredo Martínez[8], Dominique Langin[9], Jean-Michel Oppert[7], Vladimir Stich[6], Ian Macdonald[3], Peter Arner[10], Wim H. M. Saris[11], Oluf Pedersen[4], Arne Astrup[5], Philippe Froguel[2], The NUGENOB Consortium

1 Institute of Preventive Medicine, Danish Epidemiology Science Centre, Copenhagen University Hospital, Copenhagen, Denmark, **2** CNRS UPRES A 8090, Institut Biologie de Lille, Institut Pasteur de Lille, Lille, France **3** School of Biomedical Sciences, Queen's Medical Centre, University of Nottingham Medical School, Nottingham, United Kingdom, **4** Steno Diabetes Centre, Gentofte, Denmark, **5** Department of Human Nutrition, Centre for Advanced Food Research, The Royal Veterinary and Agricultural University, Copenhagen, Denmark, **6** Department of Sports Medicine, Centre of Preventive Medicine, Third Faculty of Medicine, Charles University, Praha, Czech Republic, **7** Department of Nutrition, Hôtel-Dieu Hospital, University Pierre-et-Marie Curie (Paris 6), Paris, France, **8** Department of Physiology and Nutrition, University of Navarra, Pamplona, Spain, **9** Obesity Research Unit, INSERM U586, Louis Bugnard Institute and Clinical Investigation Centre, Toulouse University Hospitals, Paul Sabatier University, Toulouse, France, **10** The Obesity Unit, Department of Medicine, Karolinska Institute, Huddinge University Hospital, Stockholm, Sweden, **11** Department of Human Biology, Nutrition, and Toxicology Research Centre NUTRIM, Maastricht University, Maastricht, Netherlands

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 8 2004, pages 1325–1326
DOI: 10.1093/bioinformatics/bth071

2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis

Jing Hua Zhao

Department of Epidemiology and Public Health, University College London, 1-19 Torrington Place, London WC1E 6BT, UK

Received on August 31, 2003; revised on November 7, 2003; accepted on November 17, 2003
Advance Access publication February 10, 2004

ASSOCIATION ANALYSIS OF UNRELATED INDIVIDUALS USING POLYMORPHIC GENETIC MARKERS

Jing Hua Zhao, Wendi Qian, University College London, UK and MRC Clinical Trials Unit, London, UK

BACKGROUND

Association analysis of unrelated individuals using multiple genetic markers are increasingly used. This could either be a marker-marker or marker-trait analysis. Haplotype phase uncertainty needs to be taken into account.

Clayton (2001) and Qin et al. (2002) have proposed heuristic EM and MCMC algorithms, but both are limited to SNPs. Here method in Clayton (2001) is extneded to multiallelic markers.

Previous global association tests using likelihoods do not give haplotype specific statistics, which are of considerable interest. We show via example they can be obtained during likelihood-based permutation tests.

METHOD AND IMPLEMENTATION

EXTENDING CLAYTON (2001)

- The new algorithm has the same feature of Clayton algorithm but considers multiple alleles when maintaining subject and haplotype lists.
- Appropriate procedure has been implemented to use results from multiple imputation as well as producing SAS programs containing the imputed data.

GLOBAL ASSOCIATION TESTS

- Likelihood-based permutation procedure is useful for producing LD-based statistics (Zhao et al. 1999)

  1. MARKER-MARKER ANALYSIS

     $\times 2$ Statistic = -2(l[assuming association]-l[linkage equilibrium])

  2. CASE-CONTROL ANALYSIS

     $\times 2$ Statistic = -2(l[case+control]-l[case]-l[control])

- Haplotype frequencies can be used for haplotype specific tests.

HAPLOTYPE SPECIFIC STATISTICS

- Simple Freeman-Turkey statistic for marker-marker analysis

$$FT = \sqrt{O} + \sqrt{O+1} - \sqrt{4E+1}$$

  O, E = haplotype counts assuming linkage disequilibrium and linkage equilibrium

- Test of proportions for case-control heterogeneity analysis

$$z = \frac{\theta_1 - \theta_2}{\sqrt{V(\theta_1 - \theta_2)}}$$

θ_1, θ_2 =haplotype frequency parameter, $V(.)$ = the variance function.

EXAMPLES

1. HLA DRB, DQA and DQB markers (25,10,15 alleles) for 94 Schizophrenic patients and 177 controls. It shows the efficiency of polymorphic markers and use of haplotype specific tests.

Table 1. Comparison of MCMC and EM estimates

Haplotype	Count	MCMC	EM	Eq	FT test	P value
22-2-12	62	11.4391	14.0193	0.4126	14.34	0.0020
4-8-1	62	11.4391	11.2545	0.5329	12.14	0.0020
9-4-1	46	8.4871	9.7786	0.3468	11.71	0.0020
1-1-7	41	7.5646	9.4095	0.2048	12.02	0.0020
6-5-3	34	6.2731	5.9737	0.2703	8.85	0.0020
8-5-3	27	4.9815	5.0698	0.1728	8.40	0.0020
14-8-2	20	3.6900	4.2366	0.1986	7.38	0.0020
6-5-2	18	3.3210	2.6979	0.3142	4.98	0.0020
17-3-13	13	2.3985	3.1291	0.0144	7.21	0.0020
10-7-6	12	2.2140	2.7675	0.0027	6.84	0.0020
21-1-9	10	1.8450	2.5830	0.0125	6.49	0.0020
18-2-14	9	1.6605	1.6605	0.0103	5.06	0.0020
3-1-7	8	1.4760	1.4760	0.0551	4.35	0.0020
9-4-4	8	1.4760	1.4760	0.067	4.26	0.0020
8-5-2	6	1.1070	1.3878	0.2009	3.35	0.0022
9-8-1	6	1.1070	<0.0001	0.5745	-2.67	N/A
12-5-4	6	1.1070	1.2915	0.0096	4.37	0.0020
16-8-2	6	1.1070	1.2915	0.0421	4.09	0.0020

Haplotype assignment by EM was unambiguous except for one individual with missing data.

Table 2. Comparison of individual haplotypes for HLA data

Haplotype	Case	Control	z-test	*P* value	Score test	*P* value
6-6-2	3.1915	0.0000	3.38	0.0003	2.95	0.0040
8-1-3	1.5957	8.2919	-3.13	0.0002	-3.11	0.0016
8-5-3	1.0638	7.2069	-3.10	0.0002	-3.05	0.0011
13-1-7	3.1915	0.2825	2.85	0.0001	2.89	0.0069
17-2-14	3.1915	0.5650	2.41	0.0008	2.45	0.0232
8-6-3	1.5957	0.0000	2.38	0.0012	2.40	0.0390
6-5-2	0.5319	3.8550	-2.27	0.0027	-2.14	0.0268
18-2-14	0.0000	2.5424	-2.20	0.0003	-2.21	0.0313
14-3-13	2.1277	0.2882	2.13	0.0023	1.38	0.3479
9-6-4	1.2395	0.0000	2.10	0.0014	1.96	0.1132
22-6-4	1.2413	0.0000	2.10	0.0008	1.96	0.1213
10-7-6	4.7872	1.6949	2.09	0.0025	2.14	0.0462
3-1-7	0.0000	2.2599	-2.08	0.0036	-2.08	0.0595
9-4-4	0.0000	2.2595	-2.08	0.0005	-2.08	0.0544

The z-statistic is comparable to score statistic, while empirical *P* values are due to different permutation procedures.

2. ALDH2 markers and 130 alcoholic patients and 133 controls. This example shows the usefulness of LD-based analysis, the effect of missing data and importance of heuristic algorithm we implemented.

Table 3. Eight ALDH2 region markers on Chromosome 12

Marker	Distance (b)	# alleles	# of missing individuals
D12S2070	> 450 000	8	251
D12S839	> 450 000	8	254
D12S821	~ 400 000	13	229
D12S1344	83 853	14	247
EXON12	0	2	261
EXON1	37 335	2	220
D12S2263	38 927	13	249
D12S1341	> 450 000	10	250

93 individuals with complete genotypes

- 1 month using only all markers by standard EM algorithm (Zhao et al. 2002)
- 6 days for 100 EM iterations using only possible haplotypes excluding two individuals with genotypes at only two loci
- 5 minutes for posterior trimming with threshold 0.001 but 8 hours with threshold 0.00001 (the new implementation)

3. 9 SNPs in APOC3/A4/A5 region from 3,012 individuals to study association with CHD and triglycerides. It shows drawbacks of heuristic algorithms and need to control for covariates.

- Log-likelihoods by Qin et al. (2002), Clayton (2001), Zhao et al.(2002) were -13,988.0, -11,607.7 and -11,521.5, respectively, suggesting increasing optimality
- 30min for Qin et al. (2002) and Clayton (2001), but 5min by Zhao et al. (2002), so the raw sorting approach is less appealing, method using sufficient statistics is desirable.
- Method of Zhao et al. (2002) also gave equilibrium likelihood

CONCLUSION

- The heuristic EM and MCMC method is able to deal with multiple multiallelic markers, but it is still difficult to use it to obtain equilibrium likelihood and sufficient statistics are necessary for large sample.
- Haplotype specific statistics an be obtained from likelihood-based implementations. They are simpler than the score statistics.

ACKNOWLEDGEMENT

We are grateful of Dr padraig Wright for providing the HLA data, Prof Alun Thomas for his unpublished manuscript and result of using possible haplotypes on ALDH2 data. We thank Dr David Clayton and Prof Jurg Ott for making their programs available. Jing hua Zhao wishes to thank Colleagues in University College and Prof Pak Sham and other Colleagues in Institute of Psychiatry for collaborative work, NIA grant to Whitehall II study (AG13196) for support.

REFERENCES

Clayton (2001). http://www-gene.cimr.cam.ac.uk/clayton/software
Qin ZS. T Niu, JS Liu (2002). Am J Hum Genet 71, 1242-7
Zhao H, Pakstis AJ, Kidd JR, Kidd KK (1999). Ann Hum Genet 63:167-179, 1999.
Zhao JH, S Lissarrague, L Essioux, PC Sham (2002). Bioinformatics 18, 1694-5

Advances and Problems…

- Advances
  - D’ and SE(D’), etc. (nontrivial since authors got wrong, and developed their MIDAS system later on)
  - Kullback-Leibler information (later AJHG paper)
  - Chromosome X data
  - Faster algorithm for multiallelic system (extension of SNPHAP)
- Problems
  - Covariates, R haplo.score, hapassoc
  - GEI interactions, R haplo.stats

The Latest?

- SAS/Genetics
- HTR
- WHAP/PLINK
- ZAPLO
  - Some ~60 programs
- But a synthesis is preferable, e.g., R
  - gap, hapassoc
  - snpMatrix, SNPassoc, GenABEL, pbatR

Journal of Statistical Software

December 2007, Volume 23, Issue 8. *http://www.jstatsoft.org/*

gap: Genetic Analysis Package

Jing Hua Zhao
MRC Epidemiology Unit

GWAS

- Hapmap
- IMPUTE and MACH
- HAPVLMC, HMM.map

Returning to chr. X

- Review of gene-counting
- Modification

ELSEVIER

Computer Methods and Programs in Biomedicine 70 (2003) 1–9

Computer Methods and Programs in Biomedicine

www.elsevier.com/locate/cmpb

Generic number systems and haplotype analysis

Jing Hua Zhao *, Pak Chung Sham

Section of Genetic Epidemiology and Biostatistics, Division of Psychological Medicine, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF, UK

Received 5 June 2001; received in revised form 21 September 2001; accepted 25 September 2001

Table 1
Genotype counts for biallelic markers

Marker 1	Marker 2		
	1/1	1/2	2/2
1/1	n_0	n_1	n_2
1/2	n_3	n_4	n_5
2/2	n_6	n_7	n_8

Table 2
Genotypic probabilities for two biallelic markers

Marker 1	Marker 2		
	1/1	1/2	2/2
1/1	h_{11}^2	$2h_{11}h_{12}$	h_{12}^2
1/2	$2h_{21}h_{11}$	$2(h_{21}h_{12}+h_{22}h_{11})$	$2h_{22}h_{12}$
2/2	h_{21}^2	$2h_{21}h_{22}$	h_{22}^2

Now Steps for X Data

- Run GENECOUNTING
  - gcx HTR2c.inp HTR2c.gc
- Run awk to extract assignment
  - awk –f HTR2c.awk HTR2c.gc > HTR2c.gco
  - HTR2c.awk has
    - /\[1\]|\[2\]/{gsub(/\[|\]/,""); print;}
- Create indicator variables for haplotypes
  - infile id chr snp1-snp4 p hid using 4snps.gco
  - tab hid, gen(h)
  - savasas using snps4.sas7bdat,replace

We Have Two Choices

- Stata with probability weighting
  - Limited success with output of parameters
- PROC SURVEYREG of SAS
  - It has ODS system such that estimates can be stacked for all six models
- How about single-locus analysis?
  - We can use a dummy variable and go through the same procedure, effectively an allellic analysis

Any Solution from R?

- Possibly with library survey

Another Example …

LRRK2 gene G2019S mutation and SNPs [haplotypes] in subtypes of Parkinson's disease

Biswanath Patra1, Azemat J. Parsian2, Brad A. Racette3, Jing Hua Zhao4, Joel S. Perlmutter3, Abbas Parsian*

1) Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center at Dallas,

Dallas, TX 2) Department of Radiation Oncology, University of Arkansas for Medical Sciences, Little Rock,

AR; 3) Department of Neurology, Washington University School of Medicine, St. Louis, MO;

4) MRC Epidemiology Unit, Strangeways Research Laboratory, Cambridge, UK

Reviewer #1:

The major limitation of this paper is the genetic association study since the study sample size does not appear to be sufficiently powered to detect the effect size differences between cases and controls. A control size of 186 is disproportionate to the cases. The authors should provide the following information:

1.A power calculation to determine if their studied sample size is sufficent to detect the effect size differences (6-loci haplotypes).

2.Provide an explanation why and how the six SNPs were selected. Have these SNPs been examined in previous published studies? Any bioinformatic data on these SNPs?

3.The age of the cases and controls does not appear to be matched. Has this been taken into account in the analysis?

4.A comment on the potential heterogeneity of the American white population since LRRK2 mutationla frequency appear to vary across some European populations.

… the study was envisaged such that it has a good though not optimal power as compared what is available in the literature. Given findings from the study, the information regarding power would only be supplementary, esp. there are also arguments that post-hoc calculation is questionable.

Nevertheless, in the case of ordinal regression (control, sporadic, familial) ~ six-SNP haplotype analysis, (via R haplo.stats library) assuming that the global association is comparable to that of a standard chi-squared statistic with value of 20.03, df=15, for a type-I error of 0.05 and a sample size of 684 used in the final analysis there would be 83% power. As this is a rough estimate, an alternative based on the 122221 haplotype between no association (haplotype frequency 0.0145) and association (haplotype freqnency 0.00874), with type-I error 0.05, a two-sided test, roughly has 30% power (32.4% with N=759 and 29.7% with N=684) according to pwr.p.test of R package pwr.

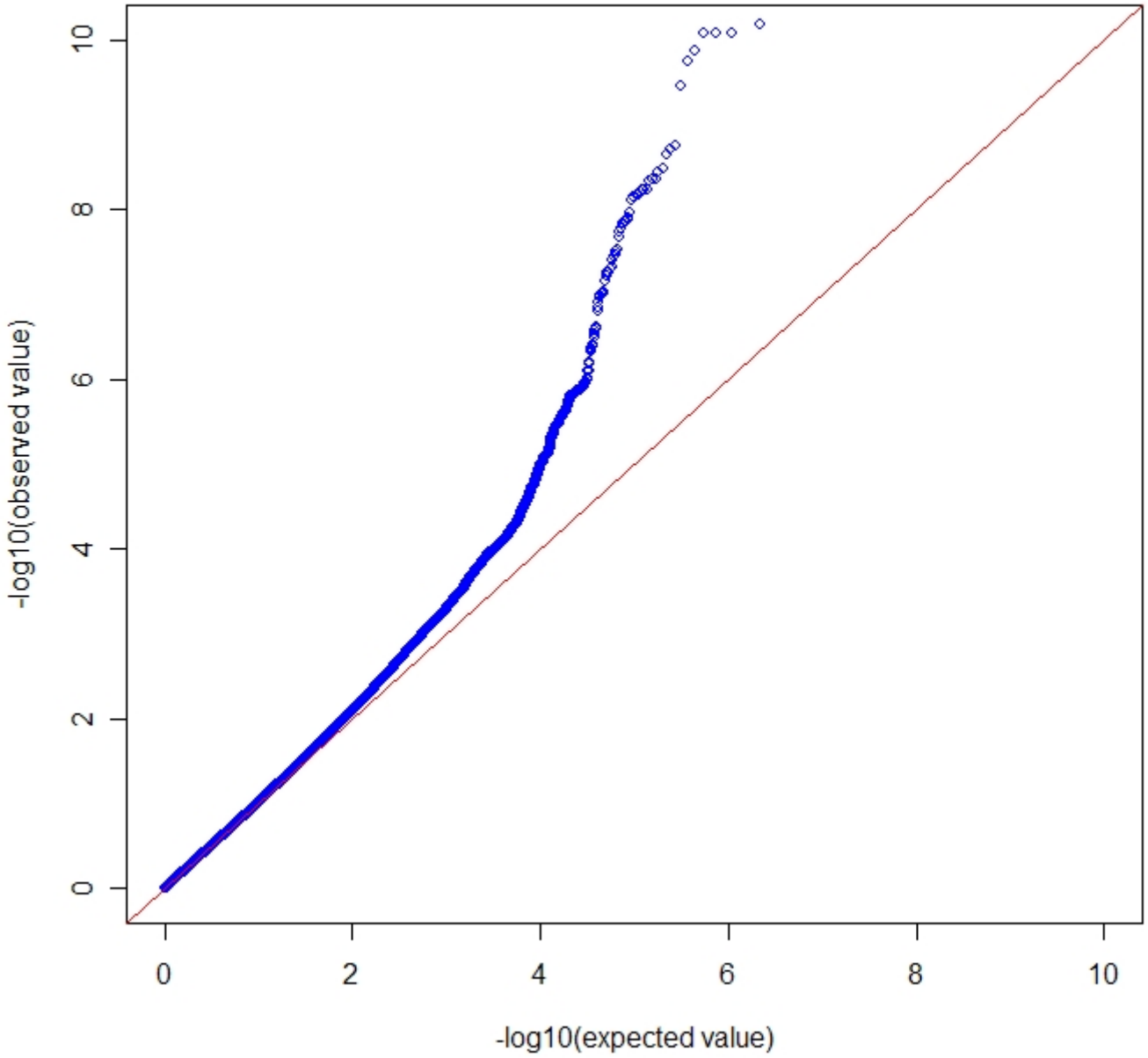
The reason to choose the six SNPs is clear from the manuscript, though not yet based on bioinformatics. The information from EMSEMBL and HapMap is consistent. In the case of HapMap which contains CEU sample (from http://www.hapmap.org/cgi-perl/gbrowse/Density_test/, type LRRK2 and search). The LRRK2 gene spans 144.3kb. Three SNPs (rs1491938, rs10784486, and rs1365763) tag 74 SNPs. The coverage is impressive although they do not cover the whole region. APART FROM rs1006151, THE OTHER TWO WERE NOT LISTED – I assume they are new relative to HapMap? Use HaploView option HapMap download for Chromosome 12 start and end positions (in Kb), Tagger, Load Includes (a column containing the six SNPs)

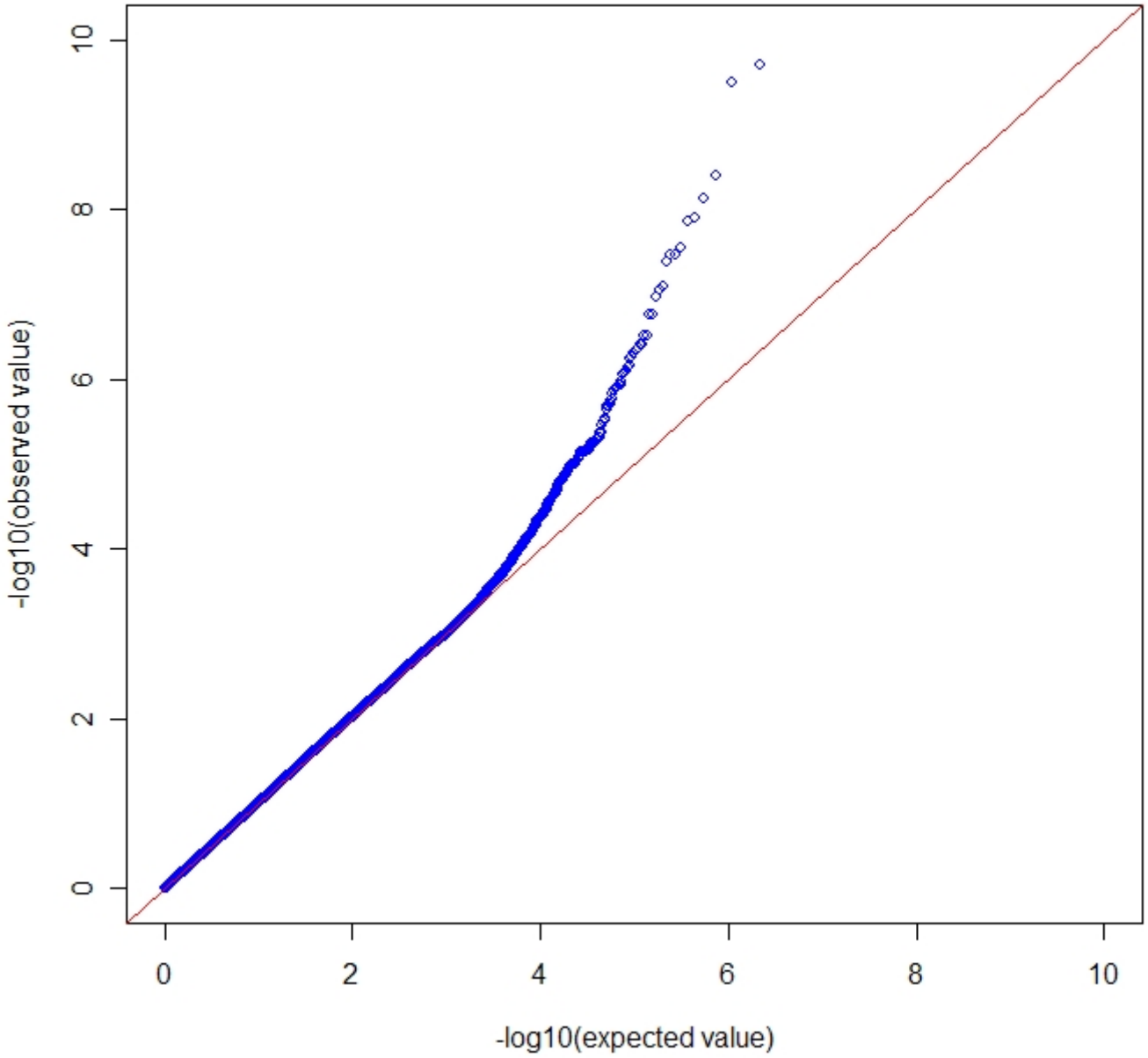
The R Trick

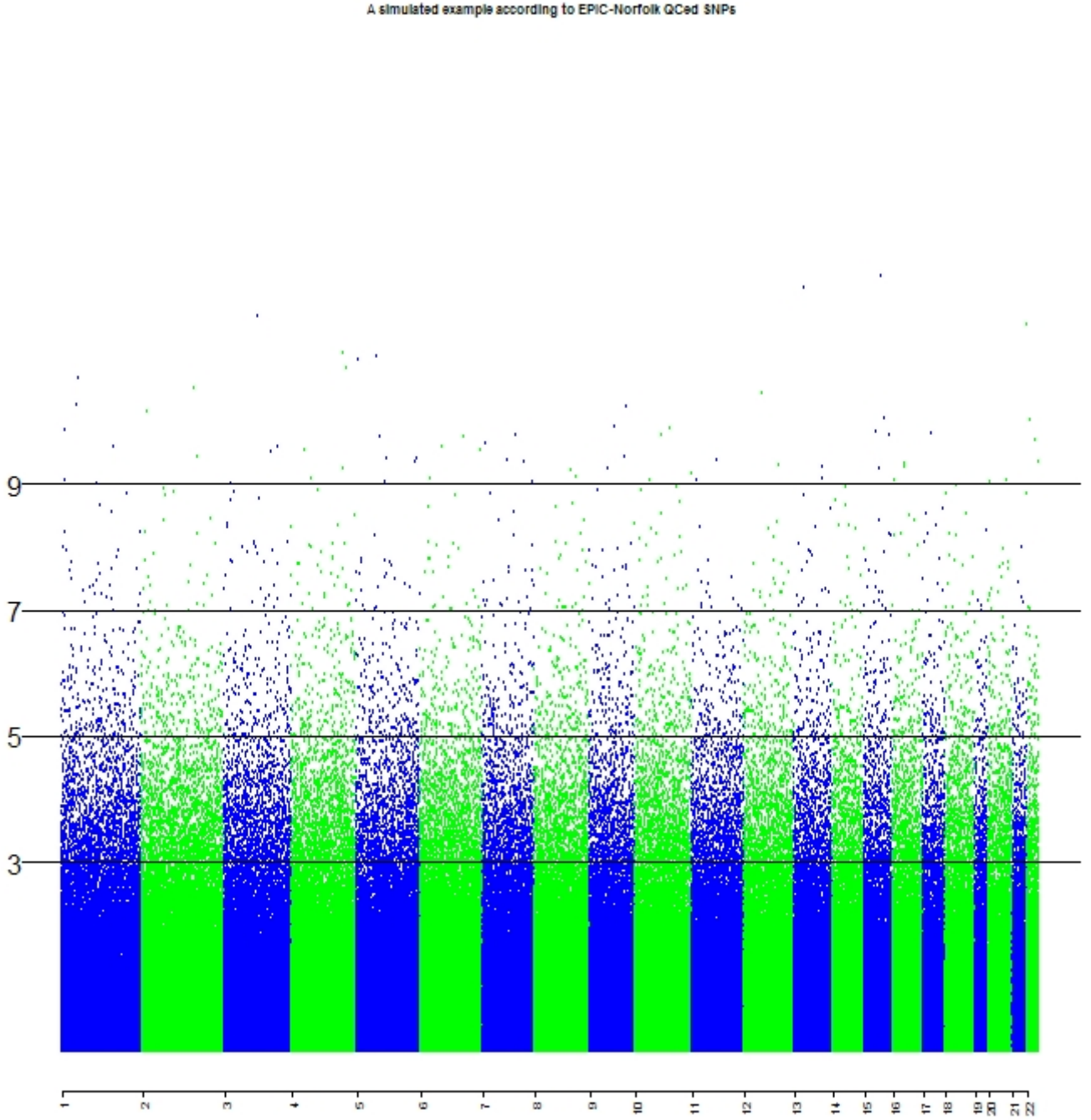
```
# 1-4-2008 MRC-Epid JHZ
# LRRK2 rely
# haplotype frequency under linkage equilibrium
p1 <- prod(c(0.15611, 0.64853,  0.84187, 0.96233, 0.58542, 0.30294))
# observed
p2 <- 0.00874
library(pwr)
p1
p2
h<-ES.h(p1,p2)
h
# had no missing data
pwr.p.test(h=h,n=759,sig.level=0.05,alternative="two.sided")
# as it is
pwr.p.test(h=h,n=684,sig.level=0.05,alternative="two.sided")
p1 <- 0.02093
p2 <- 0.00246
n1 <- 186
n2 <- 304+194
h<-ES.h(p1,p2)
h
pwr.2p2n.test(h=h,n1=n1,n2=n2,sig.level=0.05,alternative="two.sided")
n2 <- 350+225
pwr.2p2n.test(h=h,n1=n1,n2=n2,sig.level=0.05,alternative="two.sided")
```

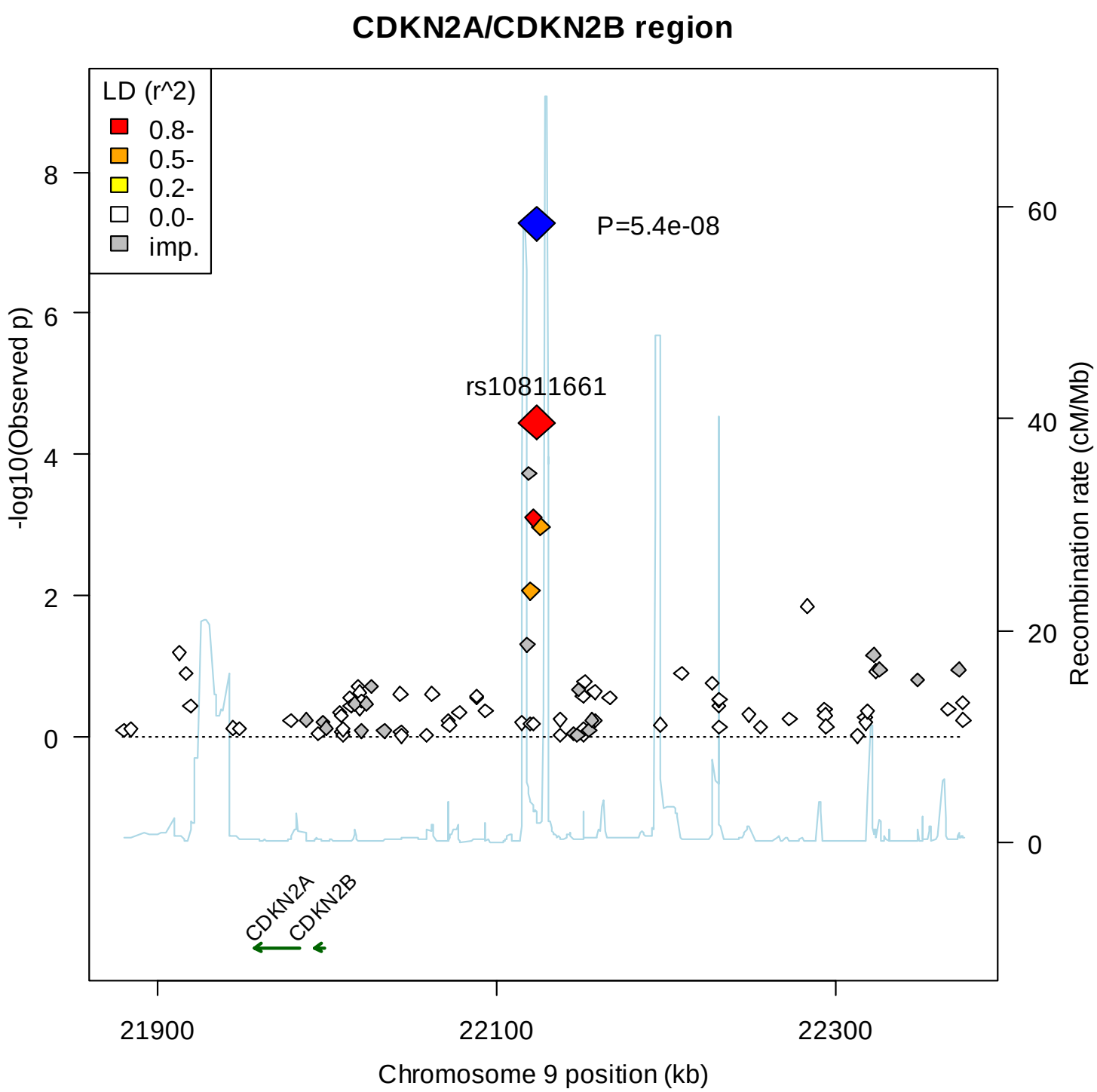
Work from 2003 Onwards

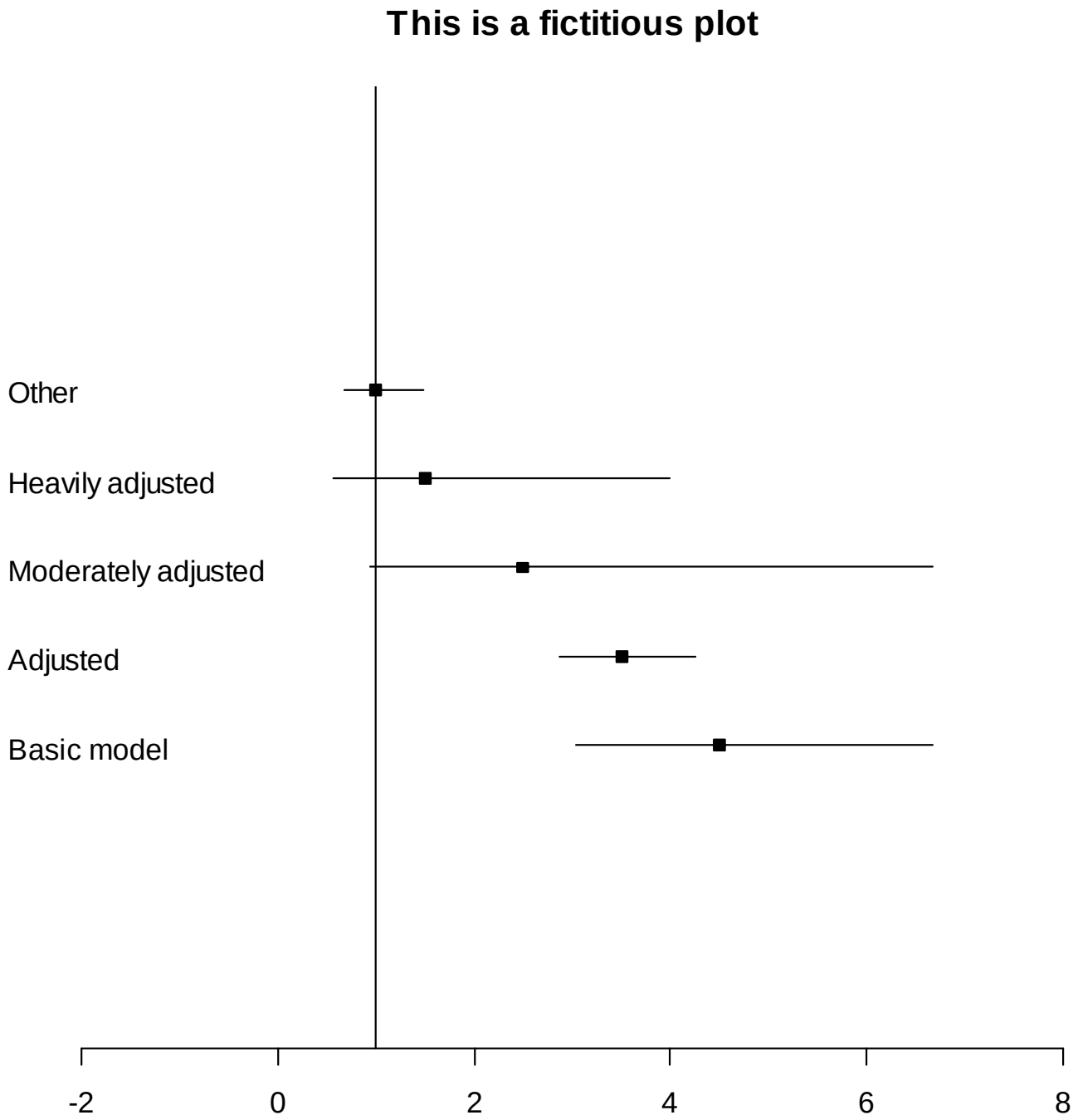
- We can do more with R …
  - A tutorial
  - Graphics: Q-Q plot, Manhanttan plot, LD plot Regional Association plot


-log10(observed value)
-log10(expected value)
0
2
4
6
8
10


-log10(observed value)
-log10(expected value)
0
2
4
6
8
10


A simulated example according to EPIC-Norfolk QCed SNPs
9
7
5
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


CDKN2A/CDKN2B region
LD (r^2)
0.8-
0.5-
0.2-
0.0-
imp.
P=5.4e-08
rs10811661
CDKN2A
CDKN2B
-log10(Observed p)
Recombination rate (cM/Mb)
Chromosome 9 position (kb)
21900
22100
22300
0
2
4
6
8
0
20
40
60


This is a fictitious plot
Other
Heavily adjusted
Moderately adjusted
Adjusted
Basic model
-2
0
2
4
6
8

More to go, but we stop here …